11 月 November
2020, 63(11): 1345 - 1357
昆　虫　学　报
ACTA ENTOMOLOGICA SINICA
http: // www. insect. org. cn
doi: 10. 16380/j. kcxb. 2020. 11. 007

基于蜜蜂球囊菌纳米孔测序数据的基因非翻译区延长、SSR 位点发掘及未注释基因和转录本鉴定

杜　宇[1,#], 付中民[1,#], 祝智威[1], 王　杰[1], 冯睿蓉[1], 王秀娜[2,3], 蒋海宾[1], 范元婵[1], 范小雪[1], 熊翠玲[1], 郑燕珍[1], 徐国钧[1], 陈大福[1], 郭　睿[1,*]

(1. 福建农林大学动物科学学院(蜂学学院), 福州 350002; 2. 福建农林大学生命科学学院, 福州 350002; 3. 福建农林大学, 福建省病原真菌与真菌毒素重点实验室, 福州 350002)

摘要:【目的】利用已获得的纳米孔长读段测序数据完善现有的蜜蜂球囊菌 *Ascosphaera apis* 参考基因组注释信息,并对未注释的新基因和新转录本进行鉴定和功能注释。【方法】基于已获得的纳米孔长读段测序数据,采用 gffcompare 软件将蜜蜂球囊菌全长转录本与参考基因组注释的转录本进行比较,进而对参考基因组注释基因的非翻译区(untranslated region, UTR)进行延长。利用 TransDecoder 软件对蜜蜂球囊菌基因的开放阅读框(open reading frame, ORF)及相应的氨基酸序列进行预测。通过 MISA 软件发掘长度在 500 bp 以上的全长转录本的 SSR 位点。通过 Blast 工具将鉴定到的新基因和新转录本比对 Nr, KOG, eggNOG, Swiss-Prot, Pfam, GO 和 KEGG 数据库进行功能注释。【结果】共对蜜蜂球囊菌的 9 481 个基因进行了 UTR 延长,其中 5′UTR 和 3′UTR 延长的基因分别有 4 744 和 4 737 个。共预测出 10 492 个完整 ORF,其中编码长度分布在 0 ~ 100 和 100 ~ 200 个氨基酸的 ORF 最多,分别占 ORF 总数的 38. 96% 和 36. 90%。共鉴定到 5 286 个 SSR,其中单核苷酸重复、二核苷酸重复、三核苷酸重复、四核苷酸重复、五核苷酸重复和六核苷酸重复的 SSR 分别为 1 870, 826, 2 398, 138, 43 和 11 个。共鉴定到 1 558 个新基因,其中有 1 556, 731, 330, 592, 1 177, 709 和 589 个新基因可分别被注释到 Nr, Swiss-Prot, Pfam, KOG, eggNOG, GO 和 KEGG 数据库。此外,还鉴定到 14 403 条新转录本,其中有 14 376, 8 524, 7 276, 7 405, 12 035, 7 891和 6 855 条新转录本可分别被注释到上述 7 个数据库。【结论】本研究利用已获得的纳米孔长读段测序数据对蜜蜂球囊菌的完整 ORF 进行了预测,对参考基因组的已注释基因进行了 UTR 延长,对未注释的 SSR 位点进行了发掘,此外还鉴定到大量未注释的新基因和新转录本,并对它们进行了功能注释。研究结果较好地完善了现有的蜜蜂球囊菌的基因组注释,为其组学和分子生物学研究的深入开展提供了基础。

关键词: 蜜蜂球囊菌; 长读段测序技术; 全长转录组; 基因组; 蜜蜂; 白垩病

中图分类号: S895. 3　　**文献标识码:** A　　**文章编号:** 0454-6296(2020)11-1345-13

Elongation of genic untranslated regions, exploration of SSR loci and identification of unannotated genes and transcripts based on the nanopore sequencing dataset of *Ascosphaera apis*

基金项目: 国家现代农业产业技术体系建设专项资金(CARS-44-KXJ7); 福建省自然科学基金项目(2018J05042); 福建省教育厅中青年教师教育科研项目(JAT170158); 福建农林大学硕士生导师团队项目(郭睿); 福建省病原真菌与真菌毒素重点实验室(福建农林大学)开放课题; 福建农林大学优秀硕士学位论文资助基金(杜宇)

作者简介: 杜宇, 男, 1994 年 8 月生, 河北唐山人, 硕士研究生, 研究方向为蜜蜂分子生物学, E-mail: m18505700830@163. com; 付中民, 男, 1972 年 1 月生, 河北唐山人, 硕士, 讲师, 研究方向为蜜蜂科学, E-mail: 369699776@qq. com

[#]共同第一作者 Authors with equal contribution

[*]通讯作者 Corresponding author, E-mail: fafu_ruiguo@126. com

收稿日期 Received: 2020-05-09; 接受日期 Accepted: 2020-06-11

DU Yu[1,#], FU Zhong-Min[1,#], ZHU Zhi-Wei[1], WANG Jie[1], FENG Rui-Rong[1], WANG Xiu-Na[2,3], JIANG Hai-Bin[1], FAN Yuan-Chan[1], FAN Xiao-Xue[1], XIONG Cui-Ling[1], ZHENG Yan-Zhen[1], XU Guo-Jun[1], CHEN Da-Fu[1], GUO Rui[1,*] (1. College of Animal Sciences (College of Bee Science), Fujian Agriculture and Forestry University, Fuzhou 350002, China; 2. College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China; 3. Key Laboratory of Pathogenic Fungi and Mycotoxins of Fujian Province, Fujian Agriculture and Forestry University, Fuzhou 350002, China)

Abstract:【Aim】This study aims to improve the annotation information of the current reference genome of *Ascosphaera apis* by utilizing previously gained nanopore long-read sequencing data, and to identify and perform functional annotation of unannotated novel genes and novel transcripts.【Methods】Based on the previously gained nanopore long-read sequencing data, full-length transcripts of *A. apis* were compared with transcripts annotated in the reference genome using gffcompare software to prolong untranslated regions (UTRs). The open reading frames (ORFs) of genes in *A. apis* and their corresponding amino acid sequences were predicted using TransDecoder software. MISA software was used to survey simple sequence repeat (SSR) loci within transcripts with a length above 500 bp. Based on Blast tool, novel genes and novel transcripts were aligned to the Nr, KOG, eggNOG, Swiss-Prot, Pfam, GO and KEGG databases to gain their corresponding functional annotations.【Results】Totally, UTRs of 9 481 genes in *A. apis* were prolonged, among which 4 744 and 4 737 genes were prolonged at 5′UTR and 3′UTR, respectively. In addition, 10 492 complete ORFs were predicted, among which the ORFs encoding proteins distributed in 0 – 100 aa and 100 – 200 aa in length were the most abundant, accounting for 38.96% and 36.90% of the total ORFs, respectively. A total of 5 286 SSRs were identified, and the numbers of mononucleotide repeats, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, pentanucleotide repeats and hexanucleotide repeats were 1 870, 826, 2 398, 138, 43 and 11, respectively. Besides, 1 558 novel genes were identified, among which 1 556, 731, 330, 592, 1 177, 709 and 589 were annotated to the Nr, Swiss-Prot, Pfam, KOG, eggNOG, GO and KEGG databases, respectively. Additionally, 14 403 novel transcripts were identified, among which 14 376, 8 524, 7 276, 7 405, 12 035, 7 891 and 6 855 were respectively annotated to the aforementioned seven databases.【Conclusion】By using the previously obtained nanopore long-read sequencing data, the complete ORFs of genes in *A. apis* has been predicted, the UTRs of annotated genes in reference genome have been elongated, the SSR loci have been explored, and a number of unannotated novel genes and novel transcripts have been identified and their functions annotated. These findings well improve the current genome annotation of *A. apis*, and offer a basis for further study on its omics and molecular biology.

Key words: *Ascosphaera apis*; long-read sequencing technology; full-length transcriptome; genome; honeybee; chalkbrood

蜜蜂是自然界最重要的授粉昆虫,在农业生产和生态维持方面发挥不可替代的作用(Montoya-Pfeiffer *et al*., 2020)。此外,蜜蜂生产的蜂王浆、蜂蜜、蜂胶和蜂蜡等蜂产品具有重要的经济和药用价值(Ahmad *et al*., 2020)。但作为群居性昆虫,蜜蜂易遭受细菌、真菌和病毒等病原微生物的侵袭而罹患疾病。其中,蜜蜂白垩病是一种长期困扰养蜂生产的顽疾,由蜜蜂球囊菌 *Ascopshaera apis* 侵染蜜蜂幼虫而引发(Jensen *et al*., 2013)。到目前为止,养蜂生产中对于白垩病仍缺乏有效的防治手段(陈大福等, 2017)。

Qin 等(2006)通过对蜜蜂球囊菌 0.5 – 1 A 和 A10 菌株进行 Sanger 测序,组装了蜜蜂球囊菌的基因组草图,但作者当时仅公布了基因序列信息,并没有同时公布基因功能注释信息,导致该版本的基因组长期无法被有效利用,阻碍了蜜蜂球囊菌的进一步研究。Shang 等(2016)运用二代测序技术对蜜蜂球囊菌 ARSEF 7405 菌株进行测序,重新组装和注释了 scafford 水平的蜜蜂球囊菌参考基因组(AAP 1.0),同时公布了完整的基因序列和基因功能注释

信息，为该真菌病原的组学和分子生物学研究奠定了基础。由于测序技术的限制，除人类（Audano *et al.*，2019）、小鼠 *Mus musculus*（Mouse Genome Sequencing Consortium，2009）和黑腹果蝇 *Drosophila melanogaster*（Solares *et al.*，2018）等极少数模式生物的基因组组装到染色体水平外，多数物种的基因组仅组装到 contig 或 scafford 水平，仍有较大的提升空间。近年来，以牛津纳米孔（Oxford Nanopore）长读段测序技术和 PacBio 单分子实时（single-molecule real-time，SMRT）测序技术为代表的三代测序技术逐渐兴起并快速发展。三代测序技术因具有超长读长的显著优势而能够轻松跨越重复序列，目前已成为基因组研究的利器（Lu *et al.*，2016；Nakano *et al.*，2017）。人们已利用纯三代测序或三代测序结合二代测序将人类（Pendleton *et al.*，2015）、跳镰猛蚁 *Harpegnathos saltator*（Shields *et al.*，2018）和苹果 *Malus domestica*（Daccord *et al.*，2017）等物种的基因组组装到染色体水平。但目前基于三代测序技术的基因组测序成本较高，对一些基因组较大的物种进行基因组测序成本仍然高昂；对于一些经费有限的实验室，利用三代测序技术进行基因组测序还存在较大困难。与基于三代测序技术的基因组测序相比，通过三代测序技术进行转录组测序的周期较短且成本较低（Magrini *et al.*，2018），因此利用三代全长转录组数据对现有的参考基因组注释进行完善是可行性较高的替代策略。近期，利用 PacBio SMRT 测序得到的全长转录组数据对锡兰勾虫 *Ancylostoma ceylanicum*（Magrini *et al.*，2018）和小麦 *Triticum aestivum*（Dong *et al.*，2015）基因组注释进行完善的研究已见诸报道。然而，利用基于 Nanopore 测序得到的长读段数据对基因组注释进行完善的研究报道匮乏。

为开展蜜蜂球囊菌的全长转录组研究，笔者前期已利用 Nanopore 长读段测序技术对蜜蜂球囊菌的纯化菌丝（AaM）和纯化孢子（AaS）分别进行测序，基于高质量的测序数据构建和注释了蜜蜂球囊菌的首个全长转录组（未发表数据）；并对蜜蜂球囊菌基因的可变剪切和可变腺苷酸化进行了系统鉴定和分析（未发表数据）。本研究利用已获得的高质量 Nanopore 长读段测序对现有的蜜蜂球囊菌参考基因组中已注释基因进行结构优化，对未注释的简单重复序列（simple sequence repeat，SSR）位点进行鉴定，进而对未注释的新基因和新转录本进行鉴定和功能注释，并预测完整开放阅读框（open reading frame，ORF）。研究结果可为蜜蜂球囊菌参考基因组的序列和功能注释提供重要补充，也能为其他物种的基因组完善提供思路和方法借鉴。

1　材料与方法

1.1　长读段测序数据来源

前期已通过 Oxford Nanopore 技术对来源于纯培养的蜜蜂球囊菌 AaM 和 AaS 分别进行全长转录组测序，获得了高质量的长读段测序数据，分别测得 6 321 704 和 6 259 727 条原始读段（raw reads），居中长度（N50）分别为 1 094 和 1 157 bp，平均读长分别为 992 和 1 047 bp，最大读长分别为 9 421 和 13 060 bp；分别鉴定出 9 859 和 16 795 条非冗余全长转录本，N50 分别达 1 482 和 1 658 bp，平均长度分别为 1 187 和 1 303 bp，最大长度分别为 6 472 和 6 815 bp（未发表数据）。纳米孔测序原始数据已上传 NCBI SRA 数据库，获得 BioProject 号：PRJNA645872。

1.2　基因结构优化

由于软件和数据本身的局限性，导致多数基因组的基因结构信息不够精确，需要进一步优化。为最大限度对蜜蜂球囊菌的参考基因组注释进行完善，本研究将 AaM 和 AaS 的长读段测序数据混合后采用 gffcompare 软件（http：//ccb. jhu. edu/software/stringtie/gffcompare. shtml）将鉴定到蜜蜂球囊菌的全长转录本与蜜蜂球囊菌参考基因组（AAP 1.0）注释的转录本进行比较，然后对基因组注释的基因结构信息进行优化。若在注释基因边界之外的区域有比对上的读段（mapped reads）支持，则将注释基因的非翻译区（untranslated region，UTR）向上游或下游延伸以修正注释基因的边界。

1.3　完整 ORF 的生物信息学预测

利用 TransDecoder 软件（http：//transdecoder. sourceforge. net/）基于 ORF 长度、对数似然函数值、氨基酸序列及 Pfam 数据库蛋白质结构域序列的比对等信息，从蜜蜂球囊菌 AaM 和 AaS 的长读段测序混合数据鉴定到的新转录本序列中识别可靠的潜在编码区序列（coding sequence，CDS）及其对应氨基酸序列，同时预测包含起始密码子和终止密码子的完整 ORF。

1.4　SSR 位点的鉴定及分析

MISA 软件（http：//pgrc. ipk-gatersleben. de/misa/）可以通过分析转录本序列鉴定出 8 种类型的 SSR，包括单核苷酸重复（p1）、双核苷酸重复（p2）、三核苷酸重复（p3）、四核苷酸重复（p4）、五核苷酸重复

(p5)、六核苷酸重复(p6)、混合 SSR(c 和 c*)(即两个 SSR 之间的距离小于 100 bp),其中 c 类型的 SSR 重复序列之间包含若干个碱基,而 c* 类型的 SSR 重复序列之间没有或只有一个其他碱基(Thiel *et al*., 2003)。从去冗余的蜜蜂球囊菌全长转录本中筛选长度在 500 bp 以上的全长转录本,利用 MISA 软件预测 SSR 位点,采用默认参数。

1.5 新基因和新转录本的鉴定及功能注释

通过将蜜蜂球囊菌的全长转录本与参考基因组注释的基因和转录本进行比较,鉴定现有参考基因组上未注释的新基因和新转录本。利用 Blast 工具将上述新基因和新转录本分别比对 Nr, Swiss-Prot, Pfam, KOG, eggNOG, GO 和 KEGG 数据库以获得相应的功能注释。

2 结果

2.1 蜜蜂球囊菌参考基因组已注释基因的 5′UTR 和 3′UTR 延长

共对蜜蜂球囊菌的 9 481 个基因的结构进行优化,其中 5′UTR 和 3′UTR 延长的基因分别有 4 744 和 4 737 个。部分蜜蜂球囊菌基因的结构优化信息如表 1 所示。

表 1 蜜蜂球囊菌参考基因组已注释的 10 个基因的结构优化信息概要

Table 1 Summary of structural optimization of ten annotated genes in the reference genome of *Ascosphaera apis*

基因 ID Gene ID	基因位置 Gene locus	正负链 Plus and minus strand	末端 End	优化前位置(bp) Original site	优化后位置(bp) Optimized site
Gene1789	AZGZ01000017.1:430545 – 432564	+	5′	430 678	430 545
Gene1789	AZGZ01000017.1:430545 – 432564	+	3′	432 202	432 564
Gene3514	AZGZ01000029.1:123607 – 124953	+	5′	123 882	123 607
Gene3514	AZGZ01000029.1:123607 – 124953	+	3′	124 590	124 953
Gene3789	AZGZ01000003.1:634688 – 637505	–	5′	635 879	634 688
Gene3789	AZGZ01000003.1:634688 – 637505	–	3′	637 204	637 505
Gene2170	AZGZ01000002.1:278896 – 281695	+	5′	279 003	278 896
Gene2170	AZGZ01000002.1:278896 – 281695	+	3′	281 340	281 695
Gene2541	AZGZ01000020.1:136027 – 137516	–	5′	136 267	136 027
Gene2541	AZGZ01000020.1:136027 – 137516	–	3′	137 354	137 516

2.2 蜜蜂球囊菌基因组中完整 ORF 预测

共预测出 10 492 个完整 ORF,它们编码的氨基酸序列长度分布介于 0 ~ 400 aa,其中分布在 0 ~ 100 aa 的 ORF 数量最多,为 4 088 个(占 38.96%);其次为分布在 100 ~ 200, 200 ~ 300 和 300 ~ 400 aa 的 ORF,数量分别为 3 872 个(占 36.90%), 1 525 个(占 14.53%)和 595 个(占 5.67%)(图 1)。

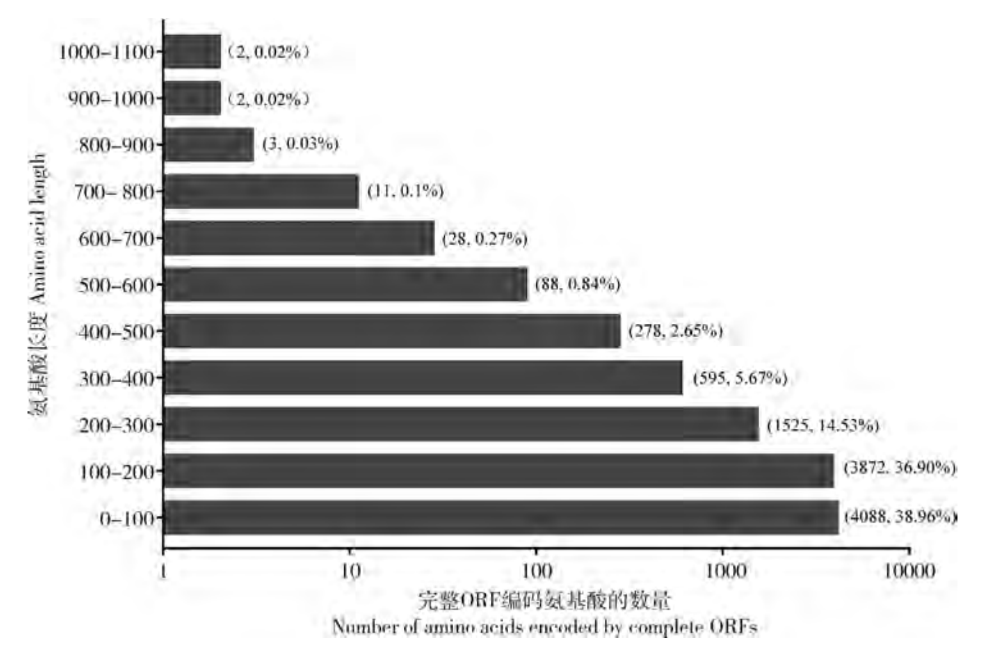


图 1 蜜蜂球囊菌基因组中完整 ORF 编码氨基酸的长度分布

Fig. 1 Length distribution of amino acids encoded by complete ORFs in the genome of *Ascosphaera apis*

2.3　蜜蜂球囊菌参考基因组未注释 SSR 位点

本研究在 24 294 167 bp 的序列中共鉴定到 5 286个 SSR 位点，含有 SSR 位点超过 1 个的基因数为 1 004 个，混合 SSR 位点有 434 个。此外，p1，p2，p3，p4，p5 和 p6 的数量分别为 1 870，826，2 398，138，43 和 11 个（表 2）。进一步分析发现，p3 类型的 SSR 密度最大，达到 83.72 个/Mb，其次为 p1，p2，c，p4，p5，c* 和 p6，分别达到 65.20，27.91，15.77，4.86，1.48，0.45 和 0.33 个/Mb（图 2）。

表 2　蜜蜂球囊菌参考基因组中 SSR 位点的 MISA 软件分析结果

Table 2　Analysis result of SSRs in the reference genome of *Ascosphaera apis* with MISA

MISA 搜索项目 MISA searching item	数目 Number
搜索基因 Searched genes	17 655
搜索基因的总序列长度 Total sequence length of searched genes (bp)	24 294 167
鉴定到的 SSR 位点 Identified SSR loci	5 286
鉴定到的 SSR 总序列长度 Total sequence length of identified SSRs (bp)	3 916
含有 1 个以上 SSR 的基因 Genes containing more than one SSR	1 004
混合 SSR Mixed SSR	434
单核苷酸重复 Mononucleotide repeats	1 870
双核苷酸重复 Dinucleotide repeats	826
三核苷酸重复 Trinucleotide repeats	2 398
四核苷酸重复 Tetranucleotide repeats	138
五核苷酸重复 Pentanucleotide repeats	43
六核苷酸重复 Hexanucleotide repeats	11

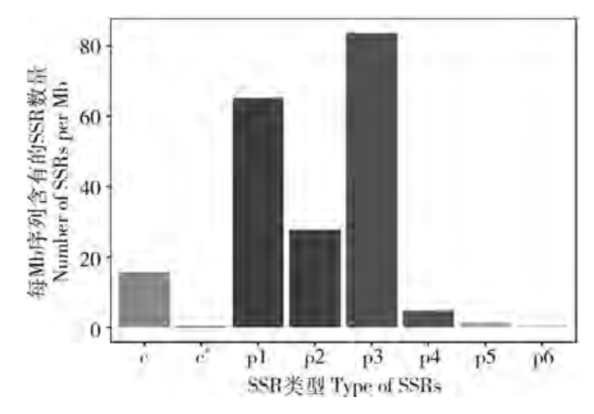


图 2　蜜蜂球囊菌参考基因组中不同类型 SSR 的密度统计

Fig. 2　Density statistics of various types of SSRs in the reference genome of *Ascosphaera apis*

p1 – p6：分别表示单核苷酸重复、二核苷酸重复、三核苷酸重复、四核苷酸重复、五核苷酸重复和六核苷酸重复的 SSR Types of SSRs with mononucleotide repeats, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, pentanucleotide repeats and hexanucleotide repeats, respectively; c, c*：混合 SSR，即两个 SSR 之间的距离小于 100 bp，其中 c 类型的 SSR 重复序列之间包含若干个碱基，而 c* 类型的 SSR 重复序列之间没有或只有一个其他碱基。Mixed SSRs, in which the distance between two SSRs is shorter than 100 bp; c indicates SSRs containing several bases, while c* indicates SSRs without other base or only with one other base.

2.4　蜜蜂球囊菌参考基因组中未注释的新基因的鉴定及功能注释

共鉴定到 1 558 个新基因，其中分别有 1 556，731，330，592，1 177，709 和 589 个新基因可分别被注释到 Nr，Swiss-Prot，Pfam，KOG，eggNOG，GO 和 KEGG 数据库。Nr 数据库中新基因注释数量最多的物种是蜜蜂球囊菌，其次为 *Polytolypa hystricis* 和伊蒙微小菌 *Emmonsia parva*（图 3：A）。新基因可注释到 KOG 数据库的 25 个功能类别，注释数量最多的是仅一般功能预测（general function prediction only），其次是翻译后修饰、蛋白质转换和分子伴侣（posttranslational modification, protein turnover, chaperones），氨基酸转运和代谢（amino acid transport and metabolism），信号转导机制（signal transduction mechanisms）以及翻译、核糖体结构和生物合成（translation, ribosomal structure and biogenesis）等（图 3：B）。此外，新基因可被注释到 eggNOG 数据库的 25 个功能类别，数量最多的为未知功能（function unknown），其次为碳水化合物转运及代谢（carbohydrate transport and metabolism），翻译后修饰、蛋白质转换和分子伴侣，细胞内移动、分泌和囊泡运输（intracellular trafficking, secretion, and vesicular transport），转录（transcription）以及翻译、核糖体结构和生物合成等（图 3：C）。

蜜蜂球囊菌的新基因还能被注释到 GO 数据库的 37 个功能条目，包括细胞组件（cell part）（347 个），细胞（cell）（340 个），细胞器（organelle）（262 个）等细胞组分相关 GO term；催化活性（catalytic activity）（328 个），结合（binding）（254 个）等分子功能相关 GO term；细胞进程（cellular process）（359 个），代谢进程（metabolism process）（340 个），单一组织进程（single-organism process）（245 个）等生物学过程相关 GO term（图 4）。

此外，上述新基因还可被注释到 KEGG 数据库的 101 条通路，包括抗生素的生物合成（biosynthesis of antibiotics）（52 个），碳代谢（carbon metabolism）（29 个），氨基酸的生物合成（biosynthesis of amino acids）（27 个），剪接体（spliceosome）（23 个），糖酵解/糖异生（glycolysis/gluconeogenesis）（20 个），细胞周期-酵母（cell cycle-yeast）（20 个），核糖体（ribosome）（18 个），RNA 转运（RNA transport）（18 个），泛素介导的蛋白水解（ubiquitin mediated

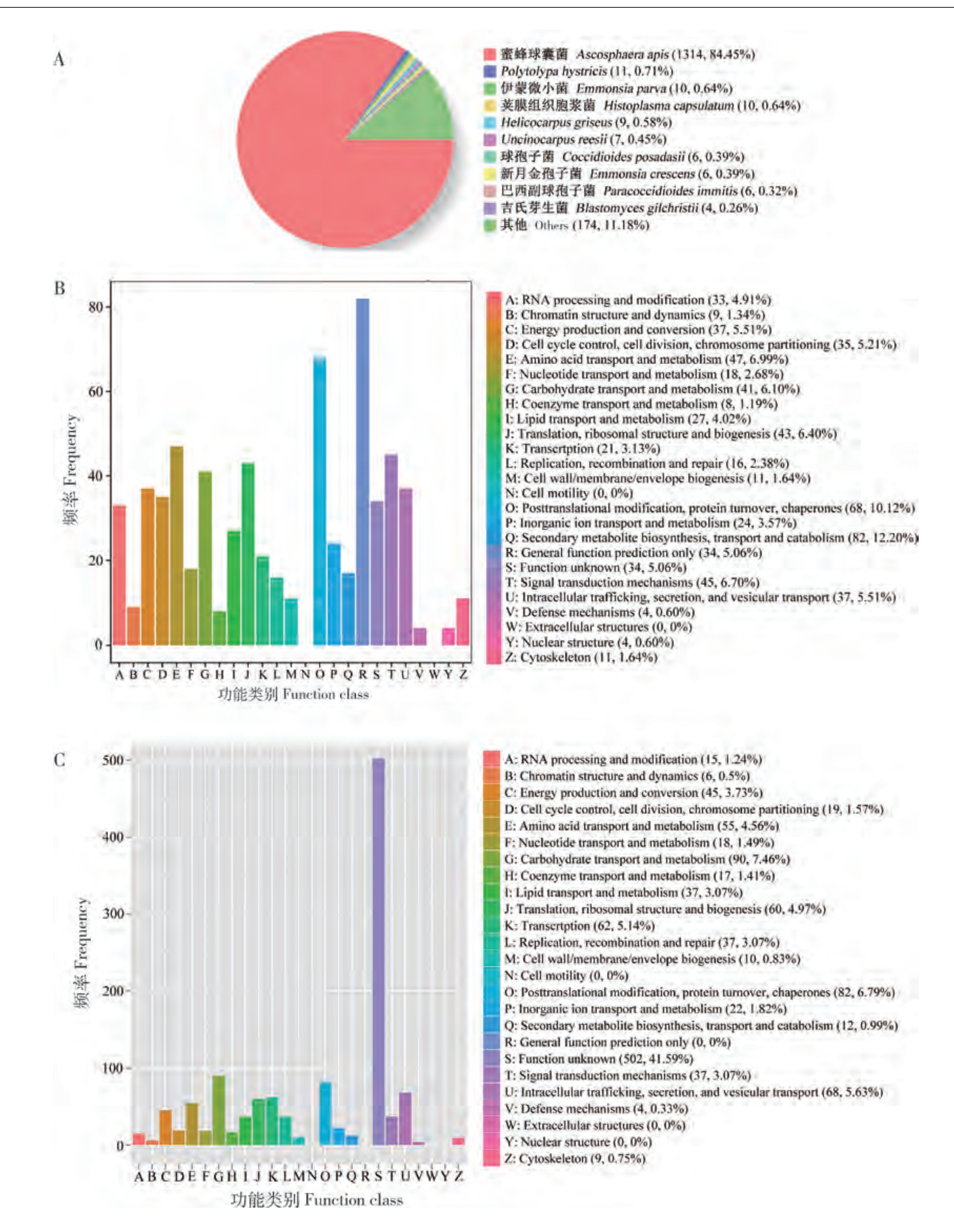


图 3 蜜蜂球囊菌参考基因组中新基因的 Nr(A)、KOG(B)和 eggNOG(C)数据库注释

Fig. 3 Annotations of novel genes in the reference genome of *Ascosphaera apis* in the Nr (A), KOG (B) and eggNOG (C) databases

proteolysis)(15 个)以及嘌呤代谢(purine metabolism)(14 个)等(图 5),条目或通路后的括号内数字代表注释的新基因占比。

2.5 蜜蜂球囊菌参考基因组中未注释的新转录本的鉴定及功能注释

共鉴定出 14 403 条新转录本,其中分别有 14 376, 8 524, 7 276, 7 405, 12 035, 7 891 和6 855 条新转录本可被分别注释到 Nr, Swiss-Prot, Pfam, KOG, eggNOG, GO 和 KEGG 数据库。Nr 数据库中新转录本注释数量最多的物种是蜜蜂球囊菌,其次为 *Polytolypa hystricis* 和 *Helicocarpus griseus*(图 6: A)。新转录本可被注释到 KOG 数据库的 25 个功

图 4　蜜蜂球囊菌参考基因组中新基因的 GO 数据库注释

Fig. 4　Annotations of novel genes in the reference genome of *Ascosphaera apis* in the GO databases

能类别，包括仅一般功能预测，翻译、核糖体结构和生物合成，翻译后修饰、蛋白质转换和分子伴侣，信号转导机制，氨基酸转运和代谢，细胞内移动、分泌和囊泡运输，能量生产和转换（energy production and conversion），RNA 加工与修饰（RNA processing and modification），未知功能以及碳水化合物转运及代谢等（图 6：B）。此外，新转录本还可被注释到 eggNOG 数据库的 25 个功能类别，包括未知功能，翻译、核糖体结构和生物合成，翻译后修饰、蛋白质转换和分子伴侣，细胞内移动、分泌和囊泡运输，碳水化合物转运及代谢，氨基酸转运和代谢，转录，能量生产和转换，脂转运及代谢（lipid transport and metabolism）以及信号转导机制等（图 6：C）。图 6 括号内数字代表注释到该条目或通路的新转录本数量和占比。

上述新转录本还能被注释到 GO 数据库的 44 个功能条目，主要涉及细胞（4 494 条），细胞组件（4 448 条），细胞器（3 356 条），细胞膜（2 332 条），大分子复合物（macromolecular complex）（1 951 条）等细胞组分相关 GO term；催化活性（3 539 条），结合（2 976 条）等分子功能相关 GO term；细胞进程（4 281 条），代谢进程（4 055 条），单一组织进程（2 584 条）等生物学过程相关 GO term（图 7）。

此外，这些新转录本还可被注释到 KEGG 数据库的 119 条通路，注释数量最多的是抗生素的生物合成（550 条），其次是核糖体（495 条），氨基酸的生物合成（284 条），碳代谢（275 条）及剪接体（253 条）等（图 8）。

3　讨论

目前，蜜蜂球囊菌的基因组尚未组装到染色体水平，其序列和功能注释信息仍需进一步优化完善。此前，笔者所在课题组利用 Illumina 测序得到的短读段数据对蜜蜂球囊菌的参考基因组注释进行完善，分别对 51 和 50 个已注释基因的 5′UTR 和 3′UTR进行延长，鉴定出 373 个新基因并对部分新基因进行了功能注释（郭睿等，2019）。Nanopore 长读段测序技术作为当前主流的三代测序技术已成功应用于人类（Lea *et al.*，2018）、大豆 *Glycine max*

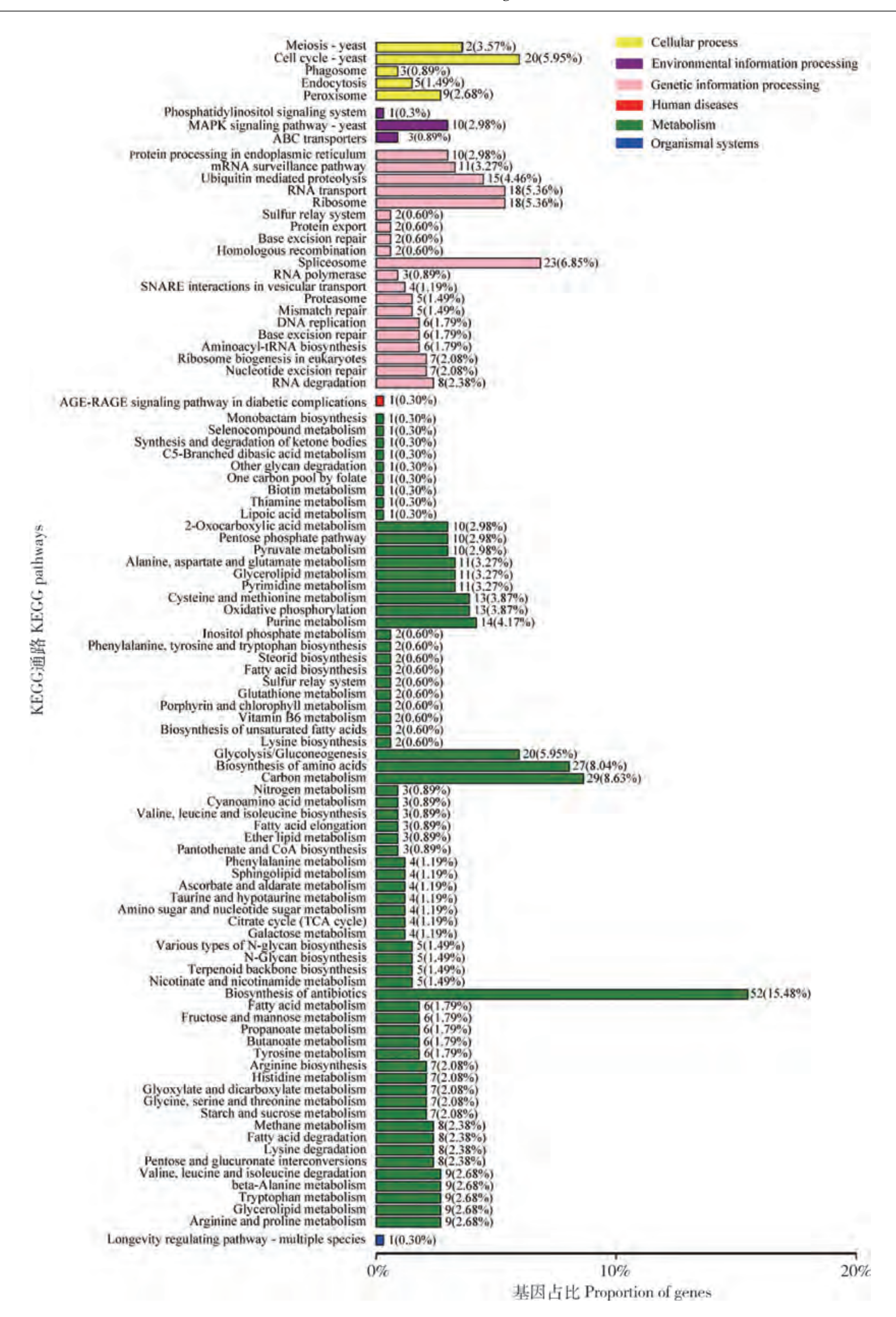


图 5 蜜蜂球囊菌参考基因组中新基因的 KEGG 数据库注释

Fig. 5 Annotations of novel genes in the reference genome of *Ascosphaera apis* in the KEGG databases

图中括号前的数字为基因数量 The numerals before brackets in the figure are the number of genes.

(Fleming *et al*., 2018)和杆状病毒(Moldován *et al*., 2018)等物种的全长转录组研究。然而对于绝大多数物种还没有基于 Nanopore 长读段测序数据完善基因组的研究报道。本研究利用前期已获得的

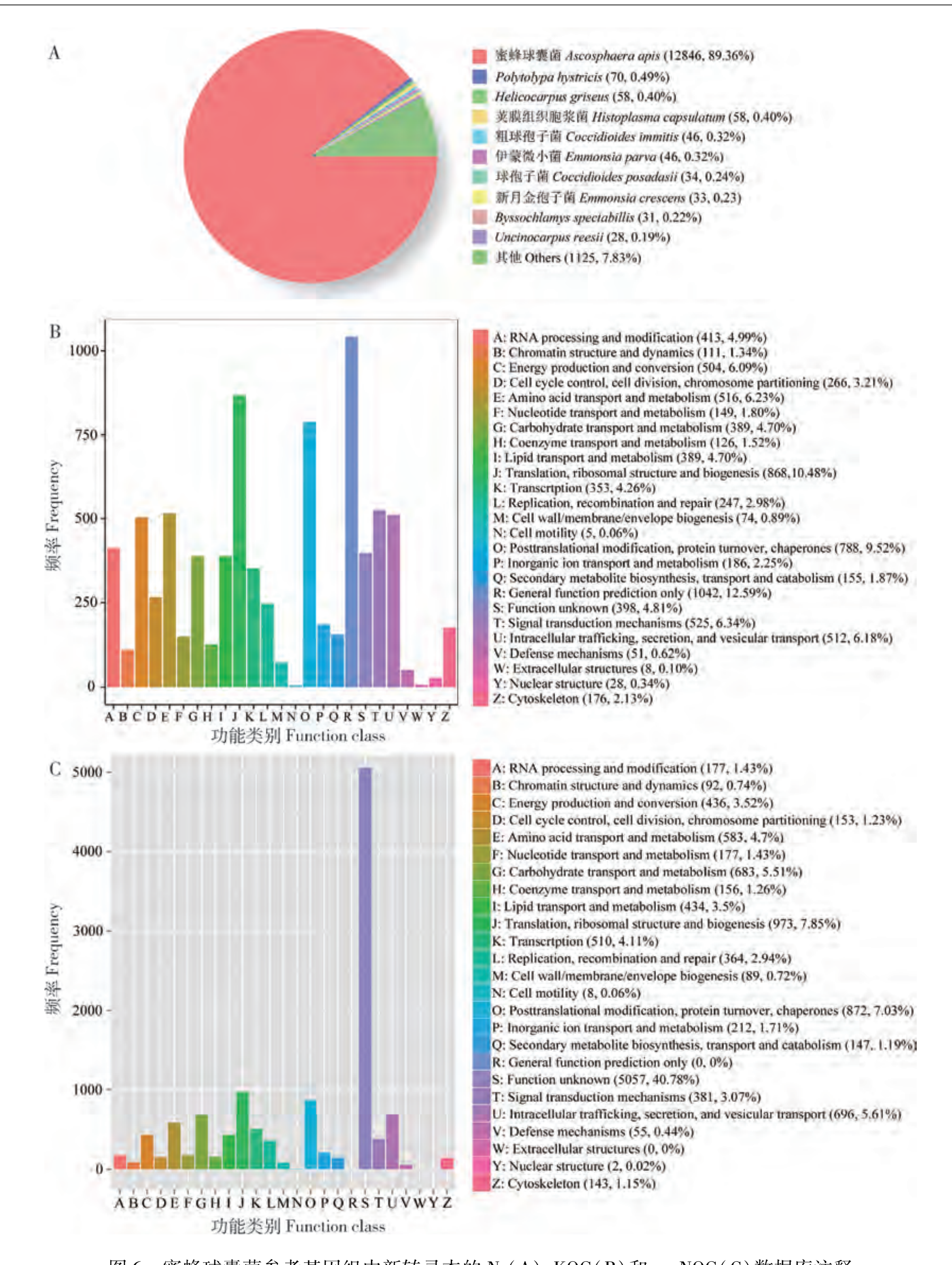


图 6　蜜蜂球囊菌参考基因组中新转录本的 Nr(A)、KOG(B)和 eggNOG(C)数据库注释

Fig. 6　Annotations of novel transcripts in the reference genome of *Ascosphaera apis* in the Nr (A), KOG (B) and eggNOG (C) databases

Nanopore 长读段测序数据对蜜蜂球囊菌的参考基因组注释进行完善，分别延长了4 744和 4 737 个已注释基因的 5′UTR 和 3′UTR，数量远多于此前基于二代测序数据延长的注释基因数量，说明 Nanopore 长读段测序技术在优化基因结构方面具有显著优势。鉴于 UTR 与真核生物的基因表达调控存在密切关系(Barrett *et al*., 2012)，本研究中蜜蜂球囊菌基因的 5′UTR 和 3′UTR 的延长对于基因表达调控的深入研究具有重要意义。此外，本研究还预测出 10 492 个完整 ORF，可为蜜蜂球囊菌基因全长序列的

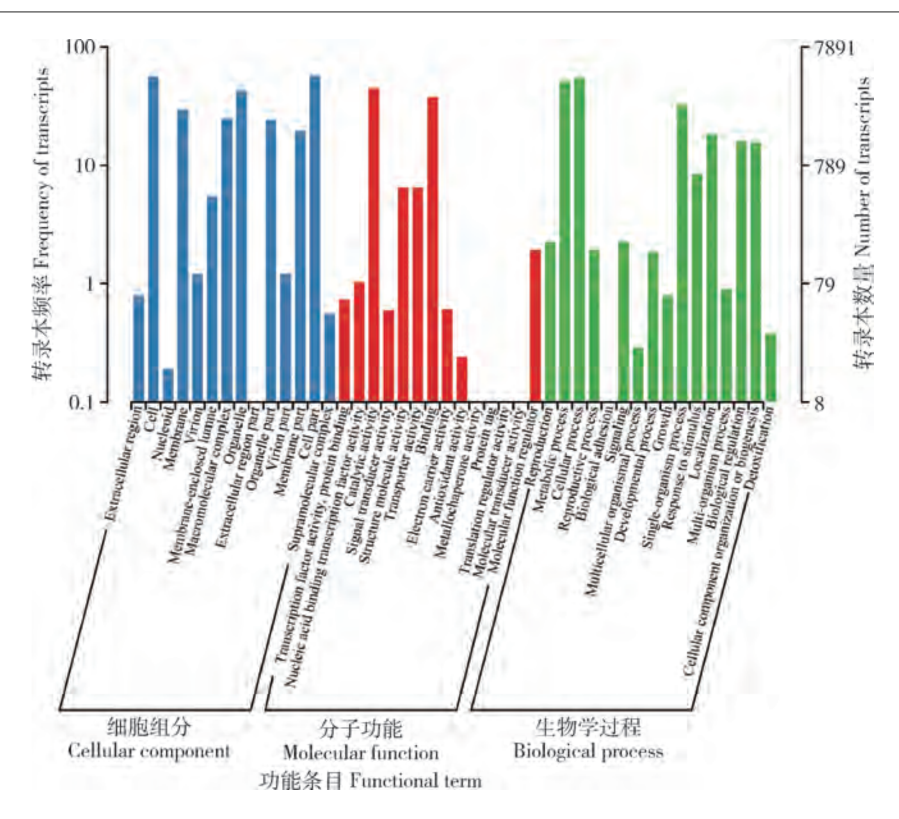


图 7　蜜蜂球囊菌参考基因组中新转录本的 GO 数据库注释

Fig. 7　Annotations of novel transcripts in the reference genome of *Ascosphaera apis* in the GO database

克隆及功能研究提供宝贵的参考信息。

第二代分子标记 SSR 是以 1 ~ 6 个核苷酸为重复单元组成的简单串联重复序列，具有实验操作易、重复性好和多态性高等优点（Jarne and Lagoda, 1996）。与传统方法相比，利用二代转录组数据开发 SSR 具有高通量的特点，使 SSR 的大规模开发成为现实（郭欢等, 2018; 黎东海和赵萍, 2019）。笔者所在课题组前期也基于 RNA-seq 数据大规模开发了中华蜜蜂 *Apis cerana cerana*（熊翠玲等, 2017）和意大利蜜蜂 *Apis mellifera ligustica*（郭睿等, 2018）的 SSR。目前，已开发和利用的蜜蜂球囊菌 SSR 较为有限。笔者所在课题组前期利用蜜蜂球囊菌的 Illumina 测序数据大规模挖掘出 7 968 个 SSR，最主要的 SSR 类型是三核苷酸重复（53.15%），其次是二核苷酸重复（32.32%）和四核苷酸重复（8.46%）（李汶东等, 2017）。本研究共鉴定到 5 286 个 SSR 位点，其中最主要的类型同样为三核苷酸重复（45.37%），其次为单核苷酸重复（35.38%）和二核苷酸重复（15.63%），表明基于三代长读段数据和二代短读段数据开发出的 SSR 类型相似，但也存在一些差异。但基于三代长读段数据开发出的 SSR 总数明显少于基于二代短读段数据开发出的 SSR 总数，究其原因，可能是前期基于二代测序数据组装得到的 unigene 总数多达 42 610 个（李汶东等, 2017），远多于蜜蜂球囊菌参考基因组包含的基因总数（6 442），这是由于二代测序得到的片段较短（不超过 300 bp），需要利用生物信息学软件对短片段进行拼接。下一步将通过毛细管电泳和荧光标记对两种测序技术开发出的 SSR 进行有效性和多态性检测，进而明确何种测序技术在大规模开发 SSR 方面更胜一筹。

前期研究中，笔者所在课题组基于蜜蜂球囊菌的 RNA-seq 数据鉴定到 373 个新基因（郭睿等, 2019）。本研究中，共鉴定到现有参考基因组未注释的 1 558 个新基因，占注释基因总数的 24.19%，说明基于 Nanopore 长读段测序数据较二代短读段测序数据在鉴定新基因方面具有显著优势。共有 1 314个新基因注释到蜜蜂球囊菌，与实际情况相符；分别有 11 和 10 个新基因注释到 *P. hystricis* 和伊蒙微小菌（图 3: A），表明上述新基因在蜜蜂球囊菌与这两个物种之间具有一定的保守性。共有 1 177个新基因可注释到 eggNOG 数据库，但注释到 Swiss-Prot, Pfam, KOG, GO 和 KEGG 数据库的新基因数量偏少，分别为 731, 330, 592, 709 和 589 个，说明这些数据库收录的蜜蜂球囊菌及近缘物种的蛋白功能注释信息较少。蜜蜂球囊菌的成熟转基因操

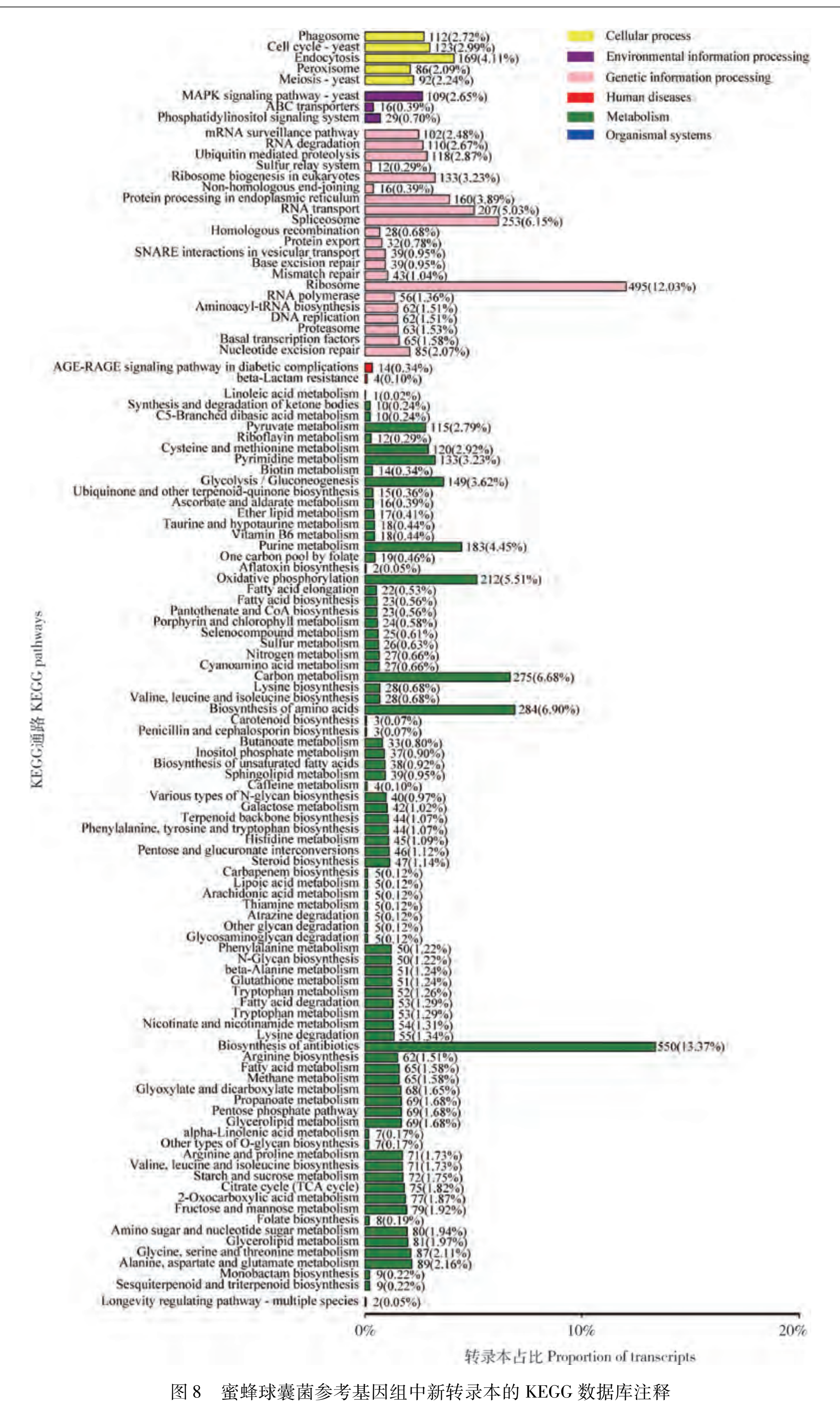


图 8　蜜蜂球囊菌参考基因组中新转录本的 KEGG 数据库注释

Fig. 8　Annotations of novel transcripts in the reference genome of *Ascosphaera apis* in the KEGG database

图中括号前的数字为转录本数量。The numerals before brackets in the figure are the number of transcripts.

作技术体系迄今尚未建立，导致蜜蜂球囊菌的基因功能研究严重滞后。近期，Tauber 等（2019）通过体外转录合成 β-葡聚糖合成蛋白编码基因以及 Ras 家族编码基因双链 RNA（dsRNA）并处理蜜蜂球囊菌，结果显示上述 dsRNA 可能在蜜蜂球囊菌孢子萌发初期被吸收，相关转录本受到抑制，孢子萌发率也相应降低。该研究为蜜蜂球囊菌的基因功能研究提供了思路借鉴。现有的蜜蜂球囊菌参考基因组注释的转录本数量为 6 442 条，本研究鉴定到 14 403 条新转录本，高于注释转录本的数量，说明由于二代测序产生的短读段的限制，蜜蜂球囊菌和其他物种的大量转录本有待挖掘，Nanopore 长读段测序技术在新转录本的鉴定方面大有作为。这些鉴定出的未注释的全长转录本可为基因全长序列克隆及功能研究提供可靠的数据基础。新转录本注释数量最多的物种同样是蜜蜂球囊菌，与现实情况相符，分别有 70 和 58 条新转录本注释到 *P. hystricis* 和 *H. griseus*（图 6：A），与新基因的注释情况略有差异。此外，分别有 14 376，8 524，7 276，7 405，12 035，7 891 和6 855条新转录本可被分别注释到 Nr，Swiss-Prot，Pfam，KOG，eggNOG，GO 和 KEGG 数据库，这些信息可进一步完善蜜蜂球囊菌的参考基因组注释。

综上所述，本研究利用高质量的 Nanopore 长读段测序数据对现有的蜜蜂球囊菌参考基因组的序列和功能注释进行了完善，为相关组学及分子生物学研究的深入开展提供了重要的参考信息，也为其他物种的基因组完善提供了方法借鉴。

参考文献（References）

Ahmad S, Campos MG, Fratini F, Altaye SZ, Li J, 2020. New insights into the biological and pharmaceutical properties of royal jelly. *Int. J. Mol. Sci.*, 21(2): 382.

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, Warren WC, Magrini V, McGrath SD, Li YI, Wilson RK, Eichler EE, 2019. Characterizing the major structural variant alleles of the human genome. *Cell*, 176(3): 663 – 675.

Barrett LW, Fletcher S, Wilton SD, 2012. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell. Mol. Life Sci.*, 69(21): 3613 – 3634.

Chen DF, Guo R, Xiong CL, Liang Q, Zheng YZ, Xu XJ, Zhang ZN, Huang ZJ, Zhang L, Wang HQ, Xie YL, Tong XY, 2017. Transcriptome of *Apis cerana cerana* larval gut under the stress of *Ascosphaera apis*. *Sci. Agric. Sin.*, 50(13): 2614 – 2623. ［陈大福，郭睿，熊翠玲，梁勤，郑燕珍，徐细建，张曌楠，黄枳腱，张璐，王鸿权，解彦玲，童新宇，2017. 中华蜜蜂幼虫肠道响应球囊菌早期胁迫的转录组学. 中国农业科学，50(13): 2614 – 2623］

Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, van de Geest H, Bianco L, Micheletti D, Velasco R, Di Pierro EA, Gouzy J, Rees DJG, Guérif P, Muranty H, Durel CE, Laurens F, Lespinasse Y, Gaillard S, Aubourg S, Quesneville H, Weigel D, van de Weg E, Troggio M, Bucher E, 2017. High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.*, 49(7): 1099 – 1106.

Dong L, Liu H, Zhang J, Yang S, Kong G, Chu JS, Chen N, Wang D, 2015. Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. *BMC Genomics*, 16(1): 1039.

Fleming MB, Patterson EL, Reeves PA, Richards CM, Gaines TA, Walters C, 2018. Exploring the fate of mRNA in aging seeds: protection, destruction, or slow decay? *J. Exp. Bot.*, 69(18): 4309 – 4321.

Guo H, Wang G, Zhang ST, Huang M, 2018. Development of SSR primers for *Simulium* (*Eusimulium*) *angustipes* (Diptera: Simuliidae) based on RNA-seq dataset. *Acta Entomol. Sin.*, 61(7): 815 – 824. ［郭欢，王刚，张树田，黄敏，2018. 基于 RNA-seq 数据的窄足真蚋 SSR 分子标记开发. 昆虫学报，61(7): 815 – 824］

Guo R, Chen HZ, Tong XY, Xiong CL, Zheng YZ, Fu ZM, Xie YL, Wang HP, Zhao HX, Chen DF, 2019. Structural optimization of annotated genes and identification of novel genes in *Ascosphaera apis*. *J. China Agric. Univ.*, 24(1): 61 – 68. ［郭睿，陈华枝，童新宇，熊翠玲，郑燕珍，付中民，解彦玲，王海朋，赵红霞，陈大福，2019. 蜜蜂球囊菌基因结构优化及新基因鉴定. 中国农业大学学报，24(1): 61 – 68］

Guo R, Chen HZ, Zhuang TY, Xiong CL, Zheng YZ, Fu ZM, Chen H, Chen DF, 2018. Exploitation of SSR markers for *Apis mellifera ligustica* based on transcriptome data. *J. Anhui Agric. Univ.*, 45(3): 404 – 408. ［郭睿，陈华枝，庄天艺，熊翠玲，郑燕珍，付中民，陈恒，陈大福，2018. 利用转录组数据开发意大利蜜蜂的 SSR 分子标记. 安徽农业大学学报，45(3): 404 – 408］

Jarne P, Lagoda PJ, 1996. Microsatellites, from molecules to populations and back. *Trends Ecol. Evol.*, 11(10): 424 – 429.

Jensen AB, Aronstein K, Flores JM, Vojvodic S, Palacio MA, Spivak M, 2013. Standard methods for fungal brood disease research. *J. Apic. Res.*, 52(1): 516 – 521.

Lea WA, Parnell SC, Wallace DP, Calvet JP, Zelenchuk LV, Alvarez NS, Ward CJ, 2018. Human-specific abnormal alternative splicing of wild-type PKD1 induces premature termination of polycystin-1. *J. Am. Soc. Nephrol.*, 29(10): 2482 – 2492.

Li DH, Zhao P, 2019. Development of microsatellite markers based on the transcriptome data of *Sclomina erinacea* (Heteroptera: Reduviidae). *Acta Entomol. Sin.*, 62(6): 694 – 702. ［黎东海，赵萍，2019. 基于转录组数据的齿缘刺猎蝽微卫星分子标记开发. 昆虫学报，62(6): 694 – 702］

Li WD, Xiong CL, Wang HQ, Hou ZX, Tong XY, Zhang L, Fu ZM, Zheng YZ, Chen DF, Guo R, 2017. Large scale development of

SSR molecular markers of *Ascosphaera apis* based on RNA-seq data. *J. Fujian Agric. For. Univ.*, 46(4): 434－438. [李汶东, 熊翠玲, 王鸿权, 侯志贤, 童新宇, 张璐, 付中民, 郑燕珍, 陈大福, 郭睿, 2017. 基于 RNA-seq 数据大规模挖掘蜜蜂球囊菌的 SSR 分子标记. 福建农林大学学报, 46(4): 434－438]

Lu HY, Giordano F, Ning ZM, 2016. Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics.*, 14(5): 265－279.

Magrini V, Gao X, Rosa BA, McGrath S, Zhang X, Hallsworth-Pepin K, Martin J, Hawdon J, Wilson RK, Mitreva M, 2018. Improving eukaryotic genome annotation using single molecule mRNA sequencing. *BMC Genomics*, 19(1): 172.

Moldován N, Tombácz D, Szücs A, Csabai Z, Balázs Z, Kis E, Molnár J, Boldogköi Z, 2018. Third-generation sequencing reveals extensive polycistronism and transcriptional overlapping in a baculovirus. *Sci. Rep.*, 8: 8604.

Montoya-Pfeiffer PM, Rodrigues RR, Alves dos Santos I, 2020. Bee pollinator functional responses and functional effects in restored tropical forests. *Ecol. Appl.*, 30(3): e02054.

Mouse Genome Sequencing Consortium, 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.*, 7(5): e1000112.

Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, Shinzato M, Minami M, Nakanishi T, Teruya K, Satou K, Hirano T, 2017. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum. Cell*, 30(3): 149－161.

Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A, Dai H, Fritz MH, Cao H, Cohain A, Deikus G, Durrett RE, Blanchard SC, Altman R, Chin CS, Guo Y, Paxinos EE, Korbel JO, Darnell RB, McCombie WR9, Kwok PY, Mason CE, Schadt EE, Bashir A, 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods*, 12(8): 780－786.

Qin X, Evans JD, Aronstein KA, Murray KD, Weinstock GM, 2006. Genome sequences of the honey bee pathogens *Paenibacillus larvae* and *Ascosphaera apis*. *Insect Mol. Biol.*, 15(5): 715－718.

Shang YF, Xiao GH, Zheng P, Cen K, Zhan S, Wang CS, 2016. Divergent and convergent evolution of fungal pathogenicity. *Genome Biol. Evol.*, 8(5): 1374－1387.

Shields EJ, Sheng L, Weiner AK, Garcia BA, Bonasio R, 2018. High-quality genome assemblies reveal long non-coding RNAs expressed in ant brains. *Cell Rep.*, 23(10): 3078－3090.

Solares EA, Chakraborty M, Miller DE2, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS, 2018. Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. *G3*, 8(10): 3143－3154.

Tauber JP, Einspanier R, Evans JD, Mcmahon DP, 2019. Co-incubation of dsRNA reduces proportion of viable spores of *Ascosphaera apis*, a honey bee fungal pathogen. *J. Apic. Res.*, 59(5): 791－799.

Thiel T, Michalek W, Varshney RK, Graner A, 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.*, 106(3): 411－422.

Xiong CL, Zhang L, Fu ZM, Wang HQ, Hou ZX, Tong XY, Li WD, Zheng YZ, Chen DF, Guo R, 2017. Large-scale development of SSR primers for *Apis cerana cerana* larvae based on its RNA-seq datasets. *J. Environ. Entomol.*, 39(1): 68－74. [熊翠玲, 张璐, 付中民, 王鸿权, 侯志贤, 童新宇, 李汶东, 郑燕珍, 陈大福, 郭睿, 2017. 基于 RNA-seq 数据大规模开发中华蜜蜂幼虫的 SSR 分子标记. 环境昆虫学报, 39(1): 68－74]

（责任编辑：马丽萍）